WATER RESOURCES RESEARCH, VOL. 46, W07516, doi:10.1029/2009WR007845, 2010

Ensemble evaluation of hydrological model hypotheses

Tobias Krueger,[1,2] Jim Freer,[3] John N. Quinton,[1] Christopher J. A. Macleod,[4] Gary S. Bilotta,[5] Richard E. Brazier,[6] Patricia Butler,[4] and Philip M. Haygarth[1]

Received 10 February 2009; revised 15 January 2010; accepted 10 March 2010; published 16 July 2010.

[1] It is demonstrated for the first time how model parameter, structural and data uncertainties can be accounted for explicitly and simultaneously within the Generalized Likelihood Uncertainty Estimation (GLUE) methodology. As an example application, 72 variants of a single soil moisture accounting store are tested as simplified hypotheses of runoff generation at six experimental grassland field-scale lysimeters through model rejection and a novel diagnostic scheme. The fields, designed as replicates, exhibit different hydrological behaviors which yield different model performances. For fields with low initial discharge levels at the beginning of events, the conceptual stores considered reach their limit of applicability. Conversely, one of the fields yielding more discharge than the others, but having larger data gaps, allows for greater flexibility in the choice of model structures. As a model learning exercise, the study points to a "leaking" of the fields not evident from previous field experiments. It is discussed how understanding observational uncertainties and incorporating these into model diagnostics can help appreciate the scale of model structural error.

Citation: Krueger, T., J. Freer, J. N. Quinton, C. J. A. Macleod, G. S. Bilotta, R. E. Brazier, P. Butler, and P. M. Haygarth (2010), Ensemble evaluation of hydrological model hypotheses, *Water Resour. Res.*, *46*, W07516, doi:10.1029/2009WR007845.

1. Introduction

[2] Hydrological models are prone to structural errors, defined, for example, by *Beven* [2005] as a combination of incorrect representations of processes, conceptual errors, processes that are not represented and implementation errors. As a consequence, several different model structures may exist for a given application, each with several different parameter sets, which may yield equally acceptable, yet imperfect, simulations when compared to the data available [*Beven and Binley*, 1992; *Neuman*, 2003; *Beven*, 2006]. Focussing on single model structures is, therefore, likely to result in modeling bias and underestimation of model uncertainty [*Neuman*, 2003].

[3] The realization of this fact has recently led to multiple model structures being considered simultaneously (ensemble simulation) in hydrological applications. Ensemble simulation studies have been undertaken in groundwater modeling [e.g., *Neuman*, 2003; *Ye et al.*, 2004; *Poeter and Anderson*, 2005], where different model structures mean mostly different models of spatially heterogeneous parameterizations. In rainfall-runoff modeling, *Shamseldin et al.* [1997] were among the first to explore simple and weighted averaging as well as neural networks as ways to combine the simulations of multiple models into a single output in some optimal way [see also *See and Abrahart* 2001]. Similar approaches to model combination, following the paradigm of a single optimal output, include: multiple-input/single-output linear transfer functions [*Shamseldin and O'Connor*, 1999]; (fuzzyfied) Bayesian inference [*See and Openshaw*, 2000]; (fuzzy) rules [*Xiong et al.*, 2001; *Abrahart and See*, 2002]; multimodel super-ensembles [*Ajami et al.*, 2006].

[4] The single model output paradigm, however, misses important information on prediction uncertainty. In contrast, *Georgakakos et al.* [2004] began to analyze the distribution of simulations within rainfall-runoff model ensembles as well as the ensemble mean. *Butts et al.* [2004] followed a similar approach in their analysis of an ensemble of structures within a common modeling framework, which they extended to the investigation of parameter and input/output data uncertainties. *Clark et al.* [2008] took a modular approach to combining the conceptual choices of four models into 79 unique structures, which they analyzed for differences and similarities.

[5] A framework to integrate all sources of uncertainty in modeling is available through Bayesian statistics. Formal Bayesian Model Averaging (BMA) was used in hydrological applications by *Vrugt et al.* [2006], *Duan et al.* [2007], *Ajami et al.* [2007] and *Vrugt and Robinson* [2007]. To overcome the usually static weighting of model structures in BMA, *Marshall et al.* [2006] introduced Hierarchical Mixtures of Experts to allow the weights of two rainfall-runoff model structures to vary dynamically depending on predicted states of a study catchment. *Hsu et al.* [2009] updated the weights of three model structures sequentially based on their performance at newly available observation time steps. An alternative to model averaging within Bayesian statistics is to

[1]Lancaster Environment Centre, Lancaster University, Lancaster, UK.
[2]Now at School of Environmental Sciences, University of East Anglia, Norwich, UK.
[3]School of Geographical Sciences, University of Bristol, Bristol, UK.
[4]Cross Institute Programme for Sustainable Soil Function, North Wyke Research, Okehampton, UK.
[5]School of Environment and Technology, University of Brighton, Brighton, UK.
[6]Department of Geography, University of Exeter, Exeter, UK.

0043-1397/10/2009WR007845

formulate a model structural error term [*Kennedy and O'Hagan*, 2001; *Vrugt et al.*, 2005; *Kuczera et al.*, 2006; *Huard and Mailhot*, 2006, 2008], although this may be problematic to define [*Beven*, 2005; *Huard and Mailhot*, 2008].

[6] An informal Bayesian framework is the Generalized Likelihood Uncertainty Estimation (GLUE) methodology, which converges to formal Bayesian inference if the required assumptions are made and likelihood measures used [*Beven*, 2006]. The possibility of multiple model structures has always been inherent in the methodology [*Beven and Binley*, 1992], although this paper is the first to explore this in an application. At the heart of GLUE is the concept of rejecting non-behavioral models and weighting the behavioral ones for ensemble simulation. Input data uncertainty can be taken into account as multiple data scenarios which are propagated through a set of models to form an extended ensemble of simulations [*Pappenberger et al.*, 2005; *Younger et al.*, 2009]. The alternative to input scenarios in a Bayesian framework is an input error term [*Kavetski et al.*, 2003; *Vrugt et al.*, 2005; *Kavetski et al.*, 2006a, 2006b; *Huard and Mailhot*, 2006; *Ajami et al.*, 2007; *Huard and Mailhot*, 2008; *Vrugt et al.*, 2008], which, again, may be difficult to estimate in practice [*Beven*, 2005; *Kavetski et al.*, 2006a]. Uncertainty in the data that models are evaluated with (output data) is usually assumed implicitly when defining model performance measures. Recent efforts, however, have made the specification of output error models more explicit [*Kennedy and O'Hagan*, 2001; *Kavetski et al.*, 2003; *Vrugt et al.*, 2003, 2005; *Beven*, 2006; *Kavetski et al.*, 2006b; *Huard and Mailhot*, 2006; *Vrugt and Robinson*, 2007; *Harmel and Smith*, 2007], although error models have rarely been justified with independent data (see *Pappenberger et al.* [2006], *Huard and Mailhot* [2008], and *Liu et al.* [2009] for exceptions).

[7] This paper demonstrates for the first time how model parameter, structural and data uncertainties can be accounted for explicitly and simultaneously within GLUE. As an example application, different model hypotheses of runoff generation are tested on a set of experimental grassland field-scale lysimeters. Following the notion of models as hypotheses of environmental systems behavior [*Beck*, 1987], this is the starting point of a downward modeling approach [*Klemeš*, 1983], i.e., one that aims first at a parsimonious description of the dynamics reflected in the observed data and then at a disaggregation of these dynamics as a continuing learning process [*Sivapalan and Young*, 2005] in which model improvement and additional data collection are interdependent. As the first iteration in this learning process, this paper is not concerned with prediction, but with model diagnostics aiming at better process representation. Input scenarios are propagated through an ensemble of conceptual models which, accounting for parameter uncertainty, are evaluated against uncertain output data. Model rejection and diagnostics are used to learn about the hydrological behavior of the study fields. Model improvement and additional data collection are suggested for the next iteration of model development.

2. Methods

2.1. Study Site

[8] Six un-drained grassland field-scale lysimeters of the Rowden Experimental Research Platform in Devon, UK (latitude 50.7802, longitude −3.9153) were investigated for the period of 01/10/2005–31/05/2006 (fields 1, 8, 10, 11, 13 and 14; Figure 1). The fields vary in area, perimeter and slope and differ in their hydrological behavior, although the soil is classified uniformly as a clayey non-calcareous pelostagnogley of the Hallsworth Series [*Avery*, 1980], a Typic Haplaquept (USDA classification) or Dystric Gleysol (FAO classification). The fields are perceived as being predominantly rain-water fed, with deep gravel-filled interceptor drains assumed to provide hydrological isolation from upslope, and 30 cm gravel-filled interceptor drains diverting overland flow and interflow through the topsoil (0–30 cm) into measurement weirs. Based on field evidence of low saturated hydraulic conductivity of the clay sub-soil ($<10^{-10}$ m s^{-1}), *Armstrong and Garwood* [1991] suggested that seepage below 30 cm is negligible.

2.2. Data and Uncertainty Estimation

2.2.1. Rainfall

[9] Four rainfall records were available from tipping buckets (Figure 1) at 1 min (gauges 1 and 2) or 1 h (gauges 3 and 4) resolution. All records were corrected for clock drift. 1.85% of the time steps where gauge 1 was obviously blocked or where data were otherwise missing were substituted with the corresponding time steps of gauge 2 and vice versa. Six rainfall scenarios were generated. The scenarios 1 and 2 are the actual records of the gauges 1 and 2. The scenarios 3–6 were created using one of the rainfall patterns of the two 1 min gauges but adjusted in rainfall volume by one of the two 1 h gauges. This adjustment resulted in a volume bias of −45.2, −51.4, −66.6 and −71.6 mm over the study period for the scenarios 3, 4, 5 and 6, respectively, because at times where either the pattern or the volume record was zero, the scenario was zero too. This volume bias was preferred over the timing error that would have resulted from a homogeneous disaggregation, because a realistic temporal rainfall pattern was expected to be important for modeling field-scale rainfall-runoff responses at 1 min resolution. It was left to the model evaluation exercise to judge if certain scenarios were infeasible in terms of field water balances. This, unfortunately, could not be assessed beforehand due to missing discharge data. It has to be kept in mind, though, that a less realistic scenario could still interact with a less feasible model structure or set of parameters to simulate seemingly acceptable discharges.

2.2.2. Evapotranspiration

[10] An automatic weather station at gauge 3 (Figure 1) further supplied hourly wind, net radiation and temperature measurements, from which potential evapotranspiration was calculated using the *Priestley and Taylor* [1972] equation, initially without the Priestley-Taylor coefficient (though see below), and neglecting the heat flux into the ground. The hourly evapotranspiration data were disaggregated homogeneously to 1 min resolution assuming a constant potential rate over the hour. An explicit treatment of the uncertainties of this model and its input data was prevented by the absence of data to characterize these. However, in the ensemble of models proposed below, multiple equations were allowed to translate potential evapotranspiration into actual rates (see Table 3), some of which include the Priestley-Taylor coefficient that was introduced to compensate for violations of the conditions of applicability of the above model [*Thornley and*

Figure 1. Location and outline of the Rowden Experimental Research Platform. The un-drained field-scale lysimeters studied are marked with numbers, and so are the four rain gauges (tipping buckets or automatic weather stations, AWS). The top right picture shows the view from rain gauge 1 toward the south-east corner of the site and the bottom left picture shows the view toward the north-west corner from the same spot. The NEXTMap Britain orthorectified radar image *Intermap Technologies* [2007] was provided courtesy of NERC via the NERC Earth Observation Data Centre (NEODC). The large-scale map is based on data provided through EDINA UKBORDERS with the support of the ESRC and JISC and uses boundary material which is copyright of the Crown and the Post Office.

Johnson, 1990]. The degrees of freedom of this ensemble of formulations will reflect at least part of the uncertainties in potential evapotranspiration.

2.2.3. Discharge

[11] Stage was measured in the weirs at 1 min (fields 10 and 13) or 5 min (fields 1, 8, 11 and 14) resolution. All records were corrected for clock drift. Gaps (Table 1) were due to failures of the wireless data transmission system or unreliable measurements. The uncertainty in the stage-discharge relationship was estimated by field experiments. Potential additional water losses across field boundaries and around weir structures could not be assessed due to the absence of relevant data. Proportionally, the importance of such losses is likely to increase at low flows. The stage-discharge uncertainty experiments were carried out on two of the weirs at a time of no runoff from the fields. In each experiment, water was fed from a tanker into the weir box and the inflow was controlled by a valve fitted to the end of the inlet hose. The inflow was increased incrementally, and once the stage reading was observed to be stable, ten repeated measurements were taken for each stage increment. The stage was recorded for each repeat to track stage drift. Corresponding discharge

Table 1. Field Statistics[a]

	Field 1	Field 8	Field 10	Field 11	Field 13	Field 14
Missing Q time steps (%)	55	0	3	83	49	72
Quick/slow Q threshold (mm 5 min^{-1})	0.0020	0.0006	0.0010	0.0001	0.0020	0.0001
Driven time steps (%)	4	8	7	6	7	6
Non-driven quick time steps (%)	25	45	37	10	33	22
Non-driven slow time steps (%)	16	47	53	1	11	0

[a]Shown are percentage of missing discharge (Q) time steps; Q threshold (center of estimated uncertainty interval) to separate "slow" from "quick" time steps; percentage of time steps driven by rain; percentage of non-driven quick time steps; and percentage of non-driven slow time steps. Remaining percentages were missing time steps.

Table 2. Estimated Error Intervals of Individual Variables Measured in the Stage-Discharge Experiments

Method	Variable	Error Interval	Comments
	Stage h (mm)	$h \pm 2$	If stage was observed to be stable
		$h \pm 5$	If stage was observed to be unstable
Bucket	Time t (s)	$t \pm 1$	
	Volume V (ml)	$V \pm 20\ n$	n is the number of transfers from bucket to measuring cylinder; 20 ml is the accuracy of the cylinder
	Max. spill (ml)	$[V;\ V + 30\ n]$	For first measurements at 10–14 mm stage done with measuring cylinder; accounts for spill to sides of cylinder due to low flows
		$[V;\ V + 10\ n]$	For measurements done with bucket; accounts for spill during transfer of water to measuring cylinder
Flowmeter	Discharge Q (l s^{-1})	$Q \pm 0.005\ Q$	Flowmeter accuracy for velocity $v > 1.5$ ft s^{-1}
		$Q \pm 0.0075\ v^{-1}\ Q$	Flowmeter accuracy for $v \leq 1.5$ ft s^{-1}; max. 0.0075 v^{-1} was 0.04 in these experiments
	Max. spill (l s^{-1})	$[Q;Q + 0.01\ Q]$	Splashing out of weir if $Q > 5$ l s^{-1}

measurements at low flows were taken using the bucket method, i.e., volumes of water were collected in a bucket (or measuring cylinder at the lowest flows) along with the time it took to fill the vessel. For high flows, discharge was measured using an electromagnetic flowmeter fitted to the end of the inlet hose. The errors in the measured variables were estimated as min/max intervals (Table 2). Combined intervals for discharge were calculated from the component variables by interval arithmetic assuming independence of component errors. The data were used to estimate the stage-discharge uncertainty using the fuzzy rating curve approach of *Pappenberger et al.* [2006], but modified here with different assumptions and a new algorithm described in Appendix A. The resultant rating curve envelope (Figure 2) has to be interpreted as min/max discharge intervals for given stages or rectangular fuzzy numbers. Since intervals could be interpreted as bounded uniform distributions, it shall be stated explicitly here that, because of the intended use of the uncertainty estimates, it was not the aim of the method to characterize the probability distribution of discharge. At this initial stage of model development, it was expected that model error would be greater than measurement error and model simulations would not fall into the observational error bounds at all time steps. Hence more detailed information about the error structure within the bounds would not add great value to overall model diagnostics for this paper.

2.3. Ensemble of Conceptual Models

[12] The confined nature of the fields lends itself intuitively to water-balance accounting via conceptual stores as the simplest initial hypothesis of runoff generation. The potentially even simpler data-based mechanistic (DBM) approach of letting the data decide upon the model structure [e.g., *Young and Beven*, 1994], within certain bounds, was not taken up here to avoid its restrictive assumptions about data and model errors. As for a more complex model, the Richards equation [*Richards*, 1931] for unsaturated flow through porous media may be theoretically applicable at the field-scale. However, due to lack of measurements of soil hydraulic properties across the fields, a homogeneous behavior would have to be assumed, at which point the approach loses its advantage of spatially more realistic runoff generation over the lumped behavior of conceptual stores. Field-scale applications of the Richards equation seem rarely, if ever, supported by data (see review by *Vereecken et al.* [2008]). In contrast, the concept of conceptual stores (see *Kirkby* [1975] and *Jothityangkoon et al.* [2001] for a comprehensive review) could be translated into different simplified hypotheses of runoff generation in this study which were testable against available data.

[13] 72 variants of a single store were considered based on the combinations of four conceptual choices, similar to the modular approach of *Clark et al.* [2008]. With the first choice it was decided whether the store was bounded or

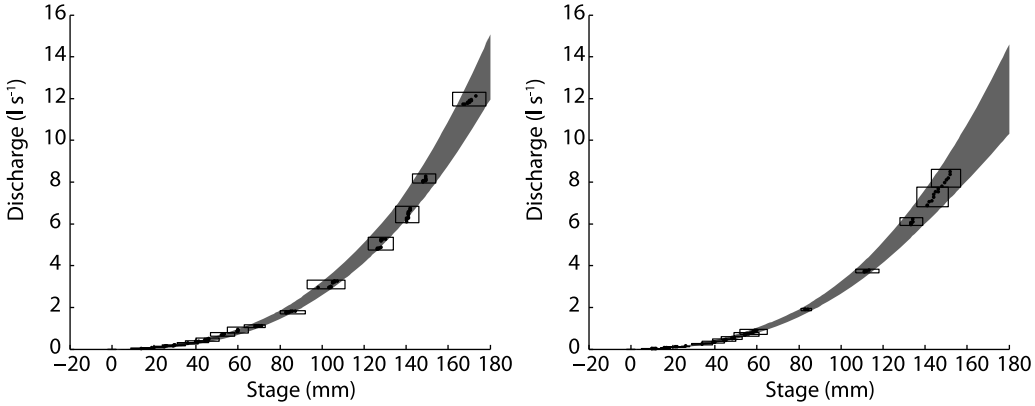

Figure 2. Rating curve envelopes derived for two weirs. Shown are the original data points, the joined data boxes for each stage increment, and the uncertainty envelope as a shaded area.

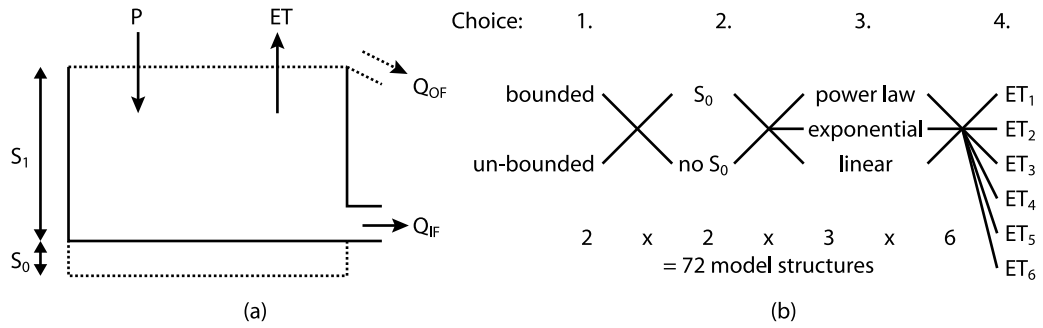

Figure 3. (a) Unified schematic of the ensemble of conceptual models. P is rainfall; ET is evapotranspiration; Q_{OF} is overland flow; Q_{IF} is interflow; S_0 is inactive storage height; and S_1 is active storage height. (b) Schematic of the 72 model structural combinations.

un-bounded (Figure 3). In the case of a bounded store, lumped saturation excess overland flow was explicitly modeled as overspill. In the case of an un-bounded store, overland flow was lumped together with interflow. With the second choice it was decided whether or not an inactive store S_0 was included which could only be accessed by evapotranspiration (Figure 3). The third choice determined the behavior of the store according to a power law, exponential or linear function. The fourth choice decided upon one of six equations to translate potential evapotranspiration into actual rates.

[14] For each of these model structures, the continuity equation (here written in discrete form)

$$\frac{\Delta S}{\Delta t} = P - ET - Q_{OF} - Q_{IF} \tag{1}$$

with storage per unit area S (mm), time step Δt (1 min or 5 min), rainfall input per unit area P (mm Δt^{-1}), evapotranspiration per unit area ET (mm Δt^{-1}), overland flow per unit area Q_{OF} (mm Δt^{-1}), and interflow per unit area Q_{IF} (mm Δt^{-1}), was solved for each time step by an explicit, forward Euler scheme. In the case of negative storage, the loss terms were adjusted to yield zero store such that the original weighting of the individual terms was preserved. Numerical errors [e.g., *Kavetski et al.*, 2006c] were minimized by using small time steps. The store was initialized to S_0 if an inactive store was included and to zero otherwise. This was realistic given that the fields did not yield any discharge in the summer months prior to the simulation period. Nevertheless, the first 41 days of the rainfall record were used to initialize the models. An independent experiment where the initial storage was sampled as an uncertain parameter confirmed the insensitivity of the model results to the initial storage after the initialization period.

[15] Overland flow was calculated as

$$Q_{OF} = \begin{cases} 0 & \text{if un-bounded} \\ S - S_1 \quad \forall \quad S > S_1 & \text{if bounded and no } S_0 \\ S - (S_1 + S_0) \quad \forall \quad S > (S_1 + S_0) & \text{if bounded and } S_0 \end{cases} \tag{2}$$

with inactive storage S_0 (mm) and active storage S_1 (mm) as identified in Figure 3.

[16] Interflow was calculated according to a power law equation

$$Q_{IF} = \begin{cases} k_p S^{m_p} \quad \forall \quad S > 0 & \text{if no } S_0 \\ k_p (S - S_0)^{m_p} \quad \forall \quad S > S_0 & \text{if } S_0 \end{cases} \tag{3a}$$

with parameters k_p (Δt^{-1}) and m_p (−); an exponential store

$$Q_{IF} = \begin{cases} k_e \exp\left(\dfrac{S}{m_e}\right) \quad \forall \quad S > 0 & \text{if no } S_0 \\ k_e \exp\left(\dfrac{S - S_0}{m_e}\right) \quad \forall \quad S > S_0 & \text{if } S_0 \end{cases} \tag{3b}$$

with parameters k_e (mm Δt^{-1}) and m_e (mm); or a linear store

$$Q_{IF} = \begin{cases} k_l S \quad \forall \quad S > 0 & \text{if no } S_0 \\ k_l (S - S_0) \quad \forall \quad S > S_0 & \text{if } S_0 \end{cases} \tag{3c}$$

with parameter k_l (Δt^{-1}). Note, the power law equation includes the linear store as a special case. The distinction, however, was made to isolate the performance of the linear store which was not possible by relying only on the power law equation due to potential parameter correlations. The same applies for the exponential store which can behave similarly to the power law store.

[17] Actual evapotranspiration was calculated by six formulations (Table 3): as the potential rate (equations (4a) and (4b)), scaled linearly with storage (equations (4c) and (4d)), or scaled as a power law function of storage (equations (4e) and (4f)). The alternative variants include an adjustment factor a to account for the potential under-estimation of the Priestley-Taylor formula in much the same way as the Priestley-Taylor coefficient. Note, the linear case was again distinguished from the power law case.

2.4. Model Diagnostics

[18] Model diagnostics shall be defined here as the analysis of model error with the aim of model improvement. This implies the need for observed data to quantify model errors and a level of spatial and temporal detail in analyzing these errors that can suggest model improvement. Previous studies have compared model performance for different periods of the hydrograph [e.g., *Freer et al.*, 1996; *Wagener et al.*,

Table 3. Six Formulations for Calculating Actual From Potential Evapotranspiration[a]

If Un-bounded and No S_0	If Bounded and No S_0	If S_0	Equation
$ET = \min(ET_{pot}, S)$	$ET = \min(ET_{pot}, S, S_1)$	$ET = \min(ET_{pot}, S, S_0)$	(4a)
$ET = \min(a\ ET_{pot}, S)$	$ET = \min(a\ ET_{pot}, S, S_1)$	$ET = \min(a\ ET_{pot}, S, S_0)$	(4b)
$ET = \min(ET_{pot}, S)$	$ET = \min(\min(\frac{S}{S_1}, 1)\ ET_{pot}, S, S_1)$	$ET = \min(\min(\frac{S}{S_0}, 1)\ ET_{pot}, S, S_0)$	(4c)
$ET = \min(a\ S\ ET_{pot}, S)$	$ET = \min(a\ \min(\frac{S}{S_1}, 1)\ ET_{pot}, S, S_1)$	$ET = \min(a\ \min(\frac{S}{S_0}, 1)\ ET_{pot}, S, S_0)$	(4d)
$ET = \min(S^b\ ET_{pot}, S)$	$ET = \min(\min(\frac{S}{S_1}, 1)^b\ ET_{pot}, S, S_1)$	$ET = \min(\min(\frac{S}{S_0}, 1)^b\ ET_{pot}, S, S_0)$	(4e)
$ET = \min(a\ S^b\ ET_{pot}, S)$	$ET = \min(a\ \min(\frac{S}{S_1}, 1)^b\ ET_{pot}, S, S_1)$	$ET = \min(a\ \min(\frac{S}{S_0}, 1)^b\ ET_{pot}, S, S_0)$	(4f)

[a]ET is actual and ET_{pot} is potential evaporation. S is the instantaneous storage height, S_0 is the inactive store, S_1 is the active store, a is an adjustment factor and b is a shape parameter. Note that equation (4c) is the same as equation (4a) for an un-bounded store with no S_0. In equations (4d–4f), a and b have a different meaning for an un-bounded store with no S_0 than for the other two cases (due to the use of S instead of a storage fraction).

2001; *Freer et al.*, 2003] or with respect to different types of data [e.g., *Freer et al.*, 2004; *Vache and McDonnell*, 2006]. Differences in parameter estimates across the hydrograph [*Wagener et al.*, 2003] or proper parameter evolution during sequential data assimilation through Kalman [e.g., *Beck*, 1987; *Vrugt et al.*, 2005] or particle [e.g., *Smith et al.*, 2008] filters have also been used to detect model inadequacies. *Wagener and Kollat* [2007] collated a suit of tools for visual model diagnostics based on Monte Carlo analysis to evaluate model identifiability, sensitivity and performance. *Clark et al.* [2008] demonstrated the link between model performance and simulated state variables (saturated area in their case) for an ensemble of model structures in order to guide model choice and improvement. The present study drew on some of the above approaches to diagnose the proposed ensemble of models within the GLUE methodology by analyzing model parameter, structural and input/output data uncertainty.

2.4.1. Model Experimental Setup

[19] For each of the 72 model structures, 100,000 parameter sets were sampled randomly from a uniform prior distribution with bounds (Table 4). Each set was run six times with one of the rainfall scenarios as model input, resulting in a total number of 43,200,000 model realizations. Every model structure and rainfall scenario was assigned the same weight, thus all realizations consisting of a model structure, a parameter set and a rainfall scenario were treated as *a priori* equally feasible hypotheses of runoff generation. The same 43,200,000 realizations were run for the six fields on a 1 min (fields 10 and 13) or 5 min (fields 1, 8, 11 and 14) time step, and were compared to the "observed" discharge uncertainty intervals.

2.4.2. Time Step-Based Performance Measure

[20] Following *Beven* [2006], the primary aim of GLUE is the rejection of non-behavioral model realizations, although it is argued that the "limits of acceptability" are difficult to define objectively. However, Beven makes a case for time step–based performance measuring that includes "effective observation errors" for the purpose of rejection, and eventually weighting and diagnosing of the remaining model realizations. For the present study, intuitive upper and lower limits of acceptability per time step would be given by the observed discharge uncertainty interval. Yet, in terms of discharge measurement error, this interval would not include potential water losses or other errors not accounted for. Nor, in terms of effective observation error, would this interval include input errors. Even though multiple rainfall scenarios were considered, none of these was error free. Hence, it was not expected that any model realization would yield simulations inside the observed discharge intervals for all time steps, and these realizations should not be rejected outright.

[21] It was thus important to define a time step–based measure of deviation D_i of simulated discharge $Q_{sim,i}$ from observed interval $Q_{obs,i}$ at time step i, which could be used for model diagnostics. This was calculated relative to the interval width, i.e., a model independent error benchmark, as

$$D_i = \frac{Q_{sim,i} - Q_{obs,i}}{\sup(Q_{obs,i}) - \inf(Q_{obs,i})} \tag{4}$$

where $\sup(Q_{obs,i})$ and $\inf(Q_{obs,i})$ are upper and lower interval bounds, respectively, and

$$Q_{sim,i} - Q_{obs,i} = \begin{cases} Q_{sim,i} - \sup(Q_{obs,i}) & \text{if } Q_{sim,i} > \sup(Q_{obs,i}) \\ 0 & \text{if } \inf(Q_{obs,i}) \le Q_{sim,i} \le \sup(Q_{obs,i}) \\ Q_{sim,i} - \inf(Q_{obs,i}) & \text{if } Q_{sim,i} < \inf(Q_{obs,i}) \end{cases} \tag{5}$$

so that $D_i = 1, 2, \ldots$ denote simulations that are 1, 2,… interval widths above the observed interval while $D_i = -1, -2, \ldots$ denote simulations that are those interval widths below. Note the "small denominator effect" in equation (4) by which, perhaps unduly, high weights were assigned to absolute deviations at low flows where interval widths were smallest (Figure 2). Especially in the case of water losses not accounted for in the estimated discharge intervals, it could be argued that the intervals at low flows should be larger. There was, therefore, a case for looking at low flow time steps separately in the next section.

2.4.3. Aggregated Performance Measures

[22] Where rigorous limits of acceptability cannot be defined and where it is computationally impossible to keep D_i for all time steps for all model realizations for the purpose of model diagnostics, a compromise has to be found aiming at a sufficiently relaxed rejection criterion that avoids the

Table 4. Model Parameter Sampling Ranges[a]

Parameter	Description	Range
S_1 (mm)	active store	0–100
S_0 (mm)	inactive store	0–60
k_p (d^{-1})	power law store parameter	0–1
m_p (−)	power law store parameter	0.1–10
k_e (mm d^{-1})	exponential store parameter	0–1
m_e (mm)	exponential store parameter	0.1–10
k_l (d^{-1})	linear store parameter	0–1
a (−)	evapotranspiration adjustment factor	0–2
b (−)	evapotranspiration shape parameter	1–10

[a]The model structures use 1–6 of these parameters depending on the conceptual choices (Figure 3 and Table 3).

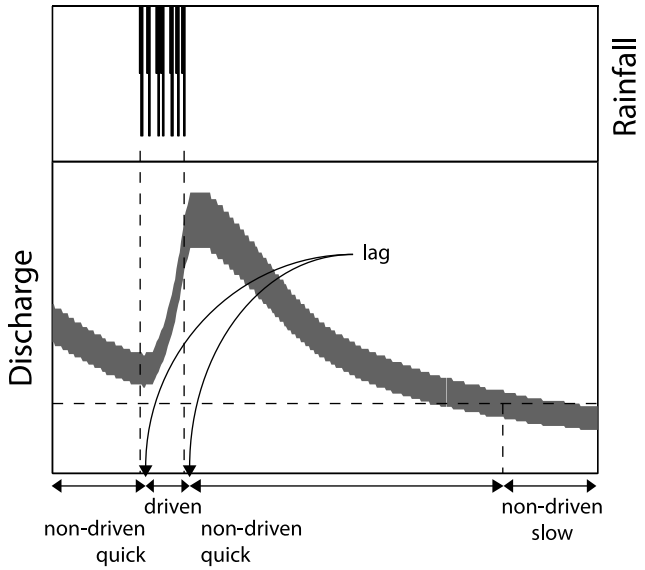

Figure 4. Idealized schematic of the hydrograph partitioning following *Boyle et al.* [2000], but modified for 1–5 min resolution. See text for the partitioning rules.

possible error of outright model rejection. To achieve this, the present study resorted to an aggregated model performance measure while keeping the time step-based information to some extent as well. For each model realization, D_i was aggregated over a number of time steps into a mean absolute D_i ($\overline{D}_{\|}$). Additionally, some information about the distribution of D_i over the particular set of time steps was retained in the form of mean negative D_i (under-predicted time steps), mean positive D_i (over-predicted time steps) and seven percentiles (min, 5th, 25th, median, 75th, 95th and max).

[23] Since aggregated performance measures can only give a balanced account of performance over a number of time steps resulting in loss of information [*Wagener et al.*, 2003], reducing this number of time steps seems crucial, even more so if the periods of aggregation can be hydrologically meaningful. This also gives rise to the possibility of deciding which are the most important periods for any given application, and model realizations can be weighted accordingly. In this study, time steps were aggregated over the three periods of the hydrograph suggested by *Boyle et al.* [2000]: periods driven by rain (performance measure $\overline{D}_{\|driven}$), non-driven high-flow ("quick") periods (performance measure $\overline{D}_{\|quick}$) and non-driven low-flow ("slow") periods (performance measure $\overline{D}_{\|slow}$). These periods are marked by dominantly different runoff generation processes, and assessing the proposed model structures on these periods means assessing their ability to describe those different processes.

[24] The hydrograph was partitioned semi-automatically following simple rules (Figure 4): The "driven" time steps were separated from the "non-driven" ones by beginning and end of rainfall, shifted by the lag between onset of rain and rise of hydrograph. If the end of rainfall fell before the hydrograph peak, the end of the driven period was moved to the peak. End points after the hydrograph peak were possible if rainfall continued beyond the peak. The "slow" time steps were separated from the "quick" ones by a discharge threshold (center of estimated uncertainty interval) defined by eye, differently for each field to take their different response characteristics into account (Table 1).

[25] To report model performance also in more familiar terms, a modified efficiency E (originally *Nash and Sutcliffe* [1970]) was calculated over all time steps as

$$E = 1 - \frac{\sum_{i=1}^{N}(Q_{sim,i} - Q_{obs,i})^2}{\sum_{i=1}^{N}(\overline{Q}_{obs} - Q_{obs,i})^2} \tag{6}$$

where $Q_{sim,i} - Q_{obs,i}$ was calculated according to equation (5). So was $\overline{Q}_{obs} - Q_{obs,i}$, but with

$$\overline{Q}_{obs} = \frac{1}{N}\sum_{i=1}^{N}\frac{\inf(Q_{obs,i}) + \sup(Q_{obs,i})}{2} \tag{7}$$

instead of $Q_{sim,i}$ in equation (5). This modification is similar to the work of *Harmel and Smith* [2007] in that observed discharge intervals are accommodated instead of "crisp" values, with the extension that $\overline{Q}_{obs} - Q_{obs,i}$ was modified here as well.

2.4.4. Model Diagnostic Scheme

[26] Sampling of the feasible parameter space for each model structure was ensured through initially wide sampling ranges (Table 4). 100,000 parameter sets were considered sufficient for the parsimonious models (1–6 parameters) used here. A model diagnostic scheme was then proposed as follows:

[27] 1. Model performance: Correlations and trade-offs between the performance measures of different periods were examined visually by drawing on elements of the "multi-criteria plot" [*Vache and McDonnell*, 2006] and the "pixel plot" [*Wagener and Kollat*, 2007], the latter to reduce the computational strain of displaying 3D correlation structures. Note, in this study, different fields may yield different correlation structures solely due to different amounts and locations of missing data.

[28] 2. Model rejection: Statistics of the D_i distribution (as well as global efficiency for comparison) were plotted against model structures and the possibility of specifying limits of acceptability to reject model structures as a whole was evaluated. The same statistics were plotted against rainfall scenarios and it was checked if certain scenarios failed in combination with any model structure based on the limits of acceptability.

[29] 3. Model weighting: Model realizations were weighted by the mean of the performance measures of the three hydrograph periods:

$$\overline{D}_{\|mean} = (\overline{D}_{\|driven} + \overline{D}_{\|quick} + \overline{D}_{\|slow})/3 \tag{8}$$

For model realizations falling within the limits of acceptability, the weights were subsequently turned into posterior GLUE likelihoods of model realization given the vector of observations $\mathbf{Q_{obs}}$ as

$$L\left(R_j(M(\boldsymbol{\theta}),\mathbf{I})|\mathbf{Q_{obs}}\right) = \frac{\overline{D}_{\|mean}^{-1} L\left(R_j(M(\boldsymbol{\theta}),\mathbf{I})\right)}{\sum_{j=1}^{J}\overline{D}_{\|mean}^{-1} L\left(R_j(M(\boldsymbol{\theta}),\mathbf{I})\right)} \tag{9}$$

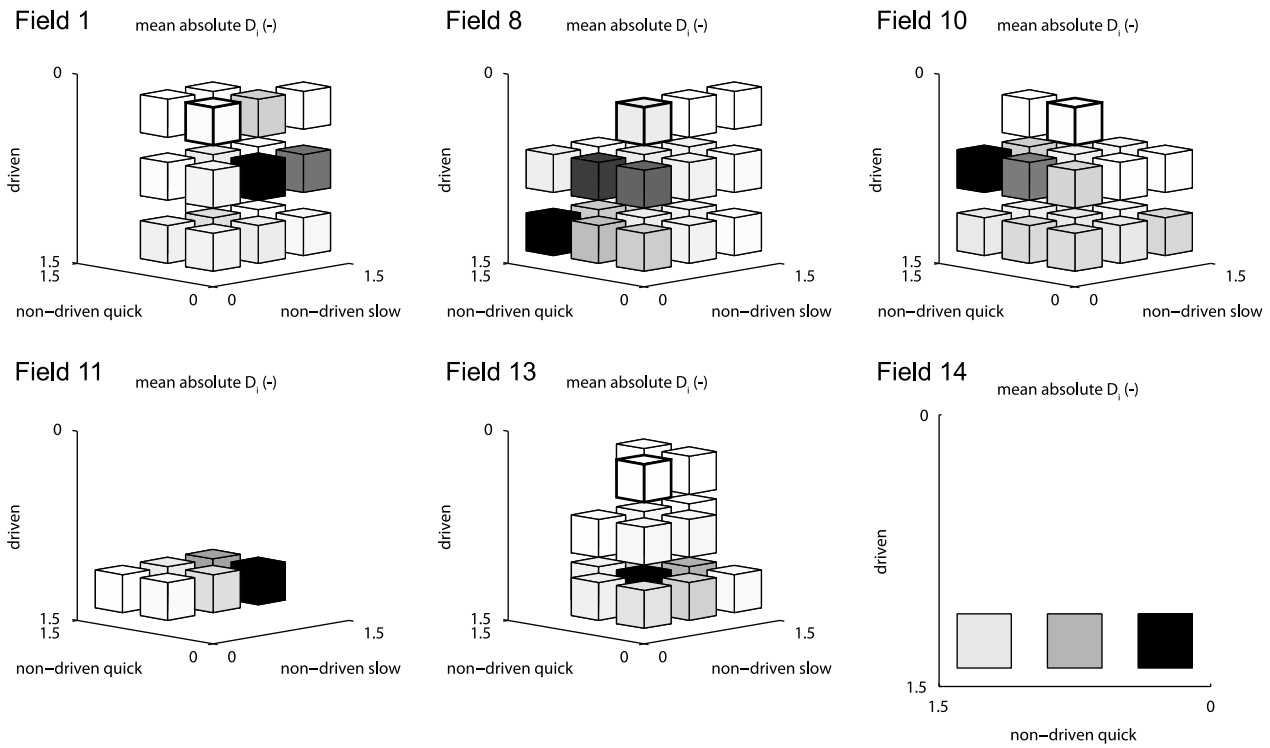

Figure 5. 3D correlation structures between mean absolute D_i ($\overline{D}_{\|}$) calculated for the driven, non-driven quick and non-driven slow periods. In each graph, the original cloud of points is discretized by equally sized cubes of 0.5 units edge length, slightly shrunk for ease of readability, with the cube of highest ranking for all three periods given a bold outline. The shading indicates the number of points within a cube, normalized by the maximum in each graph (black = maximum number of points; white = zero points). The white cubes are left out to improve readability further. Note, for field 14 non-driven slow data were missing.

with R_j being one of $j = 1,\ldots, J$ accepted model realizations, each depending on a particular model M (with parameter vector $\boldsymbol{\theta}$) and input scenario vector $\mathbf{I}$. The prior likelihood of model realization $L(R_j(M(\boldsymbol{\theta}), \mathbf{I}))$ was a constant in this study due to the uniform prior weighting of model structures, parameter sets and rainfall scenarios.

[30] 4. Model diagnostics: For the accepted ensemble of model realizations, statistics of the GLUE likelihood distribution of D_i were plotted systematically against the following hydrological variables: discharge, discharge for rising limb time steps, discharge for recession time steps, measures of antecedent wetness (discharge at onset of event and discharge sum over previous 1 min to 7 d), and season (month).

3. Results and Discussion

[31] This section follows the four items of the model diagnostic scheme proposed in the previous section.

3.1. Model Performance

[32] Figure 5 shows 3D correlation structures between $\overline{D}_{\|}$ calculated for the driven, non-driven quick and non-driven slow periods. The fields yielded different performances and correlation structures which could have been caused by different amounts of available time steps within the three periods and whether these were "easy" or "difficult" to model, but also real differences in the hydrological behavior of the fields. In this respect, field 8 was un-biased by missing time steps and field 10 was only slightly biased (Table 1). Only field 8 yielded an obvious correlation, a positive one between the driven and the non-driven quick period.

[33] Fields 1, 8, 10 and 13 yielded model realizations ranked highly for all three periods (Figure 5, cubes outlined bold). For fields 11 and 14 instead, none of the model realizations achieved such high performance with respect to the driven period. These fields had larger data gaps, although the availability of driven time steps was comparable to the other fields except field 1 (Table 1). The location of available time steps could not, therefore, explain the low performances of fields 11 and 14. Instead, these fields were marked by low initial discharge levels at the beginning of events which were indeed different to those of the other fields (compare also quick/slow discharge thresholds in Table 1). For such behavior, all models considered turned out to be rejected on the basis of a $\overline{D}_{\|}$ threshold of 0.5 for the driven time steps.

3.2. Model Rejection

[34] The model diagnostic scheme was pursued further for fields 1, 8, 10 and 13. For the model realizations where $\overline{D}_{\|} <$ 0.5 for all three hydrograph periods (Figure 5, cubes outlined bold), selected statistics of the D_i distribution were plotted against model structures (Figure 6, only the driven period is shown) and rainfall scenarios (not shown), together with global efficiency for comparison (Figure 6). Based on

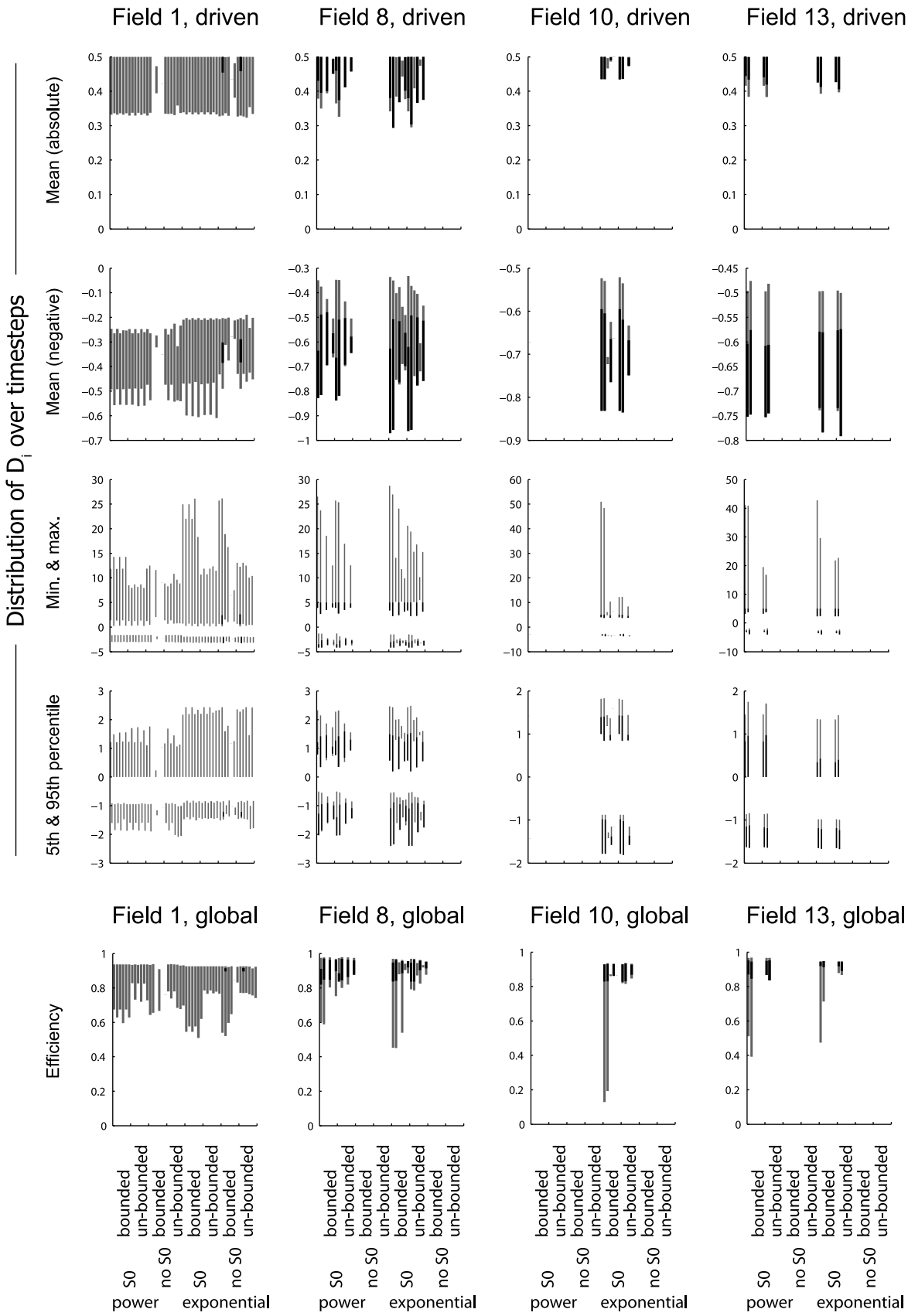
Field 1, driven
Field 8, driven
Field 10, driven
Field 13, driven
Distribution of D_i over timesteps
Mean (absolute)
Mean (negative)
Min. & max.
5th & 95th percentile
Field 1, global
Field 8, global
Field 10, global
Field 13, global
Efficiency
bounded
un-bounded
S0
no S0
power
exponential

Figure 6

the $\overline{D}_{\|}$ threshold of 0.5 applied to all hydrograph periods, the linear type of store was rejected for all fields (and is thus omitted from Figure 6), and so were model structures without the inactive store S_0 (except for field 1, see below). The power law and the exponential function yielded effectively similar storage-discharge relationships with the parameter sets accepted so far, although the power law generally resulted in more under-prediction across the fields (compare mean negative D_i statistics in Figure 6), and was thus rejected by a small margin for field 10.

[35] Overall, the ranking of model structures was similar for fields 8, 10 and 13. Field 1 was different in that model structures without the inactive store S_0 were not rejected as for the other fields. This might be explained by the fact that more discharge was observed at field 1 compared to the other fields (compare also quick/slow discharge thresholds in Table 1). This characteristic calls for a maximization of discharge in the models instead of inactive storage. It is probably also important that data from an extended dry period toward the end of the simulation were missing for field 1. The time steps of this period might have required an inactive store for modeling the threshold behavior of runoff generation during subsequent wetting up. Because of the data gaps, the analysis of field 1 is not taken further in this paper. For fields 8, 10 and 13, model performance was high, especially with respect to the modified global efficiency measure which could exceed 0.9 (Figure 6). The following analysis, therefore, delves into the more subtle issues of model performance.

[36] The choice of evapotranspiration function seemed to be more important than whether stores were bounded or un-bounded, with equations (4a) and (4b) favored (Figure 6). In fact, the storage parameter S_1 was so high in the model realizations shown here that the bounded realizations were hardly ever saturated and simulated overland flow was minimal. The bounded stores then reacted effectively as un-bounded ones. An independent investigation confirmed that modeling overland flow as overspill routed to the field outlet in one time step caused unrealistic over-predictions using 1–5 min time steps. Explicit overland flow routing would be required to account for the necessary lag and attenuation, although the concept of homogeneous generation of overland flow across the fields is itself not realistic.

[37] The maximum over-prediction was still unrealistically high for other parameter sets and model structures (see min & max statistics in Figure 6), which also resulted in low efficiency values. These extremes were obviously not picked out by the $\overline{D}_{\|}$ criterion, hence an upper limit of acceptability of $D_i \leq 5$ was applied to reject those model realizations. A symmetrical lower limit of $D_i \geq -5$ was chosen. Note, all model realizations could have been rejected using a stricter limit, were it not for the need to retain some realizations for model diagnostics. In the realm beyond the more "objective" observational error bounds, it will only be possible to scrutinize limits of acceptability further relative to future improved models. Finally, it was impossible to reject any rainfall scenario for all fields within the setup of this study, likely because of compensational effects between rainfall scenarios and model parameters which is investigated in the next section.

3.3. Model Weighting

[38] The model realizations falling within the limits of acceptability of $-5 \leq D_i \leq 5$ and meeting the $\overline{D}_{\|}$ threshold of 0.5 for all three hydrograph periods were weighted according to equation (8) with corresponding GLUE likelihoods of model realization after equation (9). Figure 7 shows the accepted ensembles of model structures and rainfall scenarios for fields 8, 10 and 13. The ensembles were generally composed of the same model structures across the fields, albeit different relative contributions and performances. The highest weights were associated with the exponential type of store and the evapotranspiration equation (4b). Rainfall scenarios, too, showed different relative contributions to the accepted ensembles and different performances across the fields.

[39] Figure 8 zooms further into the accepted model structures, exemplified for field 13. The un-bounded variants of the accepted model structures are not shown as they exhibited virtually the same correlation plot matrices as the bounded variants for the parameters other than S_1. Obvious correlations existed between k_p and m_p of the power law type of store (not shown) and between k_e and m_e of the exponential type of store (Figure 8). Correlations also existed between rainfall scenarios and the evapotranspiration adjustment factor a of equation (4b) (Figure 8). The adjustment of the ET_{pot} estimates to higher values (increasing a) for scenario/gauge 1 reflects the overall higher rainfall of this gauge. All three fields favored values of a close to or larger than 1 (see Figure 8 for field 13) leading to values of ET close to or larger than ET_{pot} whenever the store was filled sufficiently. This resulted in a total simulated evapotranspiration flux over the Water Year which was almost as high as the total simulated discharge flux (Figure 9a, shown as GLUE likelihood distribution) and only slightly less than the total estimated potential evapotranspiration flux of 499 mm a^{-1}.

[40] Figure 9b shows the GLUE likelihood distribution of the maximum simulated store S_{max} for each field which shall be called "effective pore space" here, the conceptual equivalent of soil pore space minus residual soil moisture. Note that elements of storage representing overland flow and the interceptor drains are lumped into S_{max} as well. For comparison, field data suggests a porosity of 48% for this soil type of which 23% is residual soil moisture and 41% is soil field capacity. Together with the assumed topsoil depth of 30 cm this works out at an equivalent S_{max} of 111 mm, larger than the effective pore space suggested by the model results. Even if the topsoil depth was only 20 cm, the equivalent S_{max} would be with 74 mm at the upper end of the distribution of model results (Figure 9b). The inactive store S_0 is the conceptual equivalent of soil field capacity, shown as percentage of effective pore space S_0/S_{max} and GLUE likelihood distribu-

Figure 6. Selected statistics of the D_i distribution over the driven period and global efficiency plotted against model structures for fields 1, 8, 10 and 13. Each grey bar represents the extent of model realizations for one model structure (a multidimensional parameter and rainfall scenario space itself). Where two statistics are combined into one graph, two bars share one slot of a model structure. The nested black bars represent the extent of model realizations that fell within the limits of acceptability of $-5 \leq D_i \leq 5$.

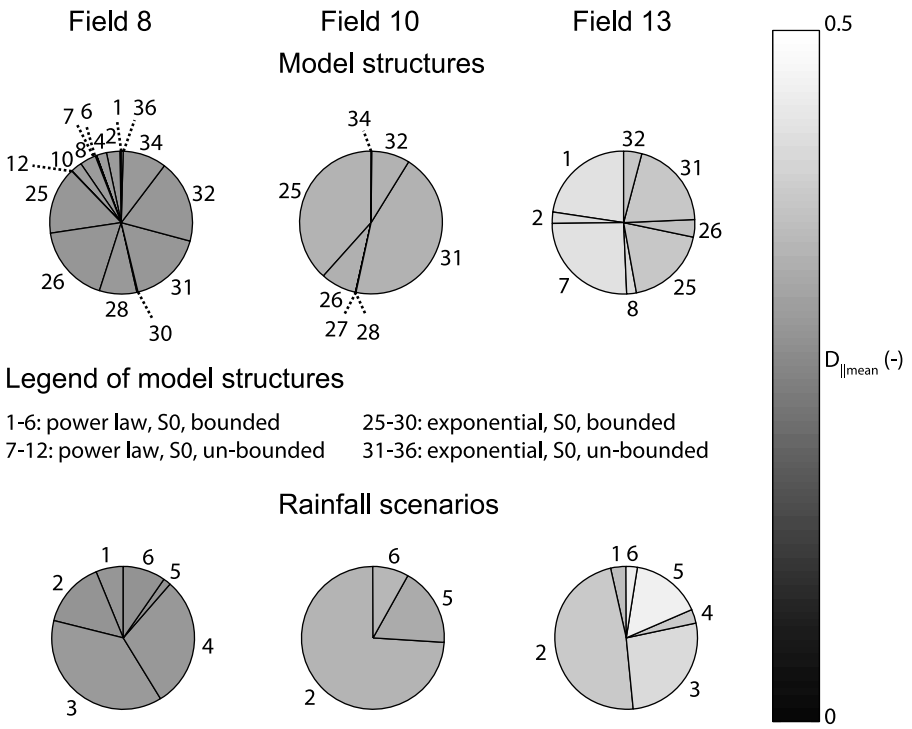

Figure 7. Accepted ensembles of model structures and rainfall scenarios for fields 8, 10 and 13 (the 48 model structures rejected for all three fields are omitted). The pie chart fractions are proportional to the relative contributions of structures/scenarios to each ensemble, the numbers are labels (see legend) and the shading is proportional to the average weighting of model realizations within each structure/scenario (the "highest" weight ($\overline{D}_{||mean}$ = 0) in black and the "lowest" weight ($\overline{D}_{||mean}$ = 0.5) in white).

tion in Figure 9c. For comparison, the field data estimates suggest a lower S_0/S_{max} equivalent of 53%.

3.4. Model Diagnostics

[41] The accepted ensembles of model realizations for fields 8, 10 and 13 were analyzed for systematic deviations between simulations and observations, i.e., deviations associated with certain flow regimes (high/low, rising/falling), certain states of antecedent wetness (formalized as discharge at onset of event and discharge sum over previous 1 min to 7 d) or season (month). The dominant systematic factors were discharge magnitude and rise/fall of the hydrograph (Figure 10). Incidentally, Figure 10 also provides a comparison of the time step–based performance measure D_i (a deviation relative to the observed discharge interval width) with absolute deviations (Q_{obs} against Q_{sim}). Since the estimated discharge interval width was a convex function of discharge (center of interval; Figure 2), the absolute deviations at low flows were inflated through D_i relative to the same absolute deviations at high flows, resulting in the dominant convex decrease of D_i (from both positive and negative values toward zero) with increasing discharge that can be seen in Figure 10. When this is understood, Figure 10 conveys a greater GLUE likelihood of over-predicting the low flows and under-predicting the high flows, and this was more pronounced during recession periods (Figure 11 gives an example). This behavior was similar across the fields, although the simulations for field 8 were generally closer to the observed intervals and the under-prediction of the rising time steps at high flows was less pronounced.

4. Conclusions

[42] This paper demonstrated how model parameter, structural and data uncertainties can be accounted for explicitly and simultaneously within the Generalized Likelihood Uncertainty Estimation (GLUE) methodology. With the inclusion of multiple model structures, the logical extension of the GLUE paradigm of testing multiple model hypotheses was realized for the first time. It was shown that discharge error estimates and, by implication, those of other evaluation data can serve as model independent benchmarks for testing model hypotheses. However, the understanding of data uncertainties will often remain incomplete, in this study particularly with respect to rainfall input. This, and the need for retaining imperfect models for diagnostic or opera-

Figure 8. Representation of the accepted rainfall scenario/parameter space of field 13 in the form of a correlation plot matrix for two variants of the exponential, bounded store with inactive store S_0: that using evapotranspiration equation (4a) on the lower left triangle and that using equation (4b) on the upper right triangle. S_1 is the active store, k_e and m_e are the parameters of the exponential store, and a is the evapotranspiration adjustment factor of equation (4b).

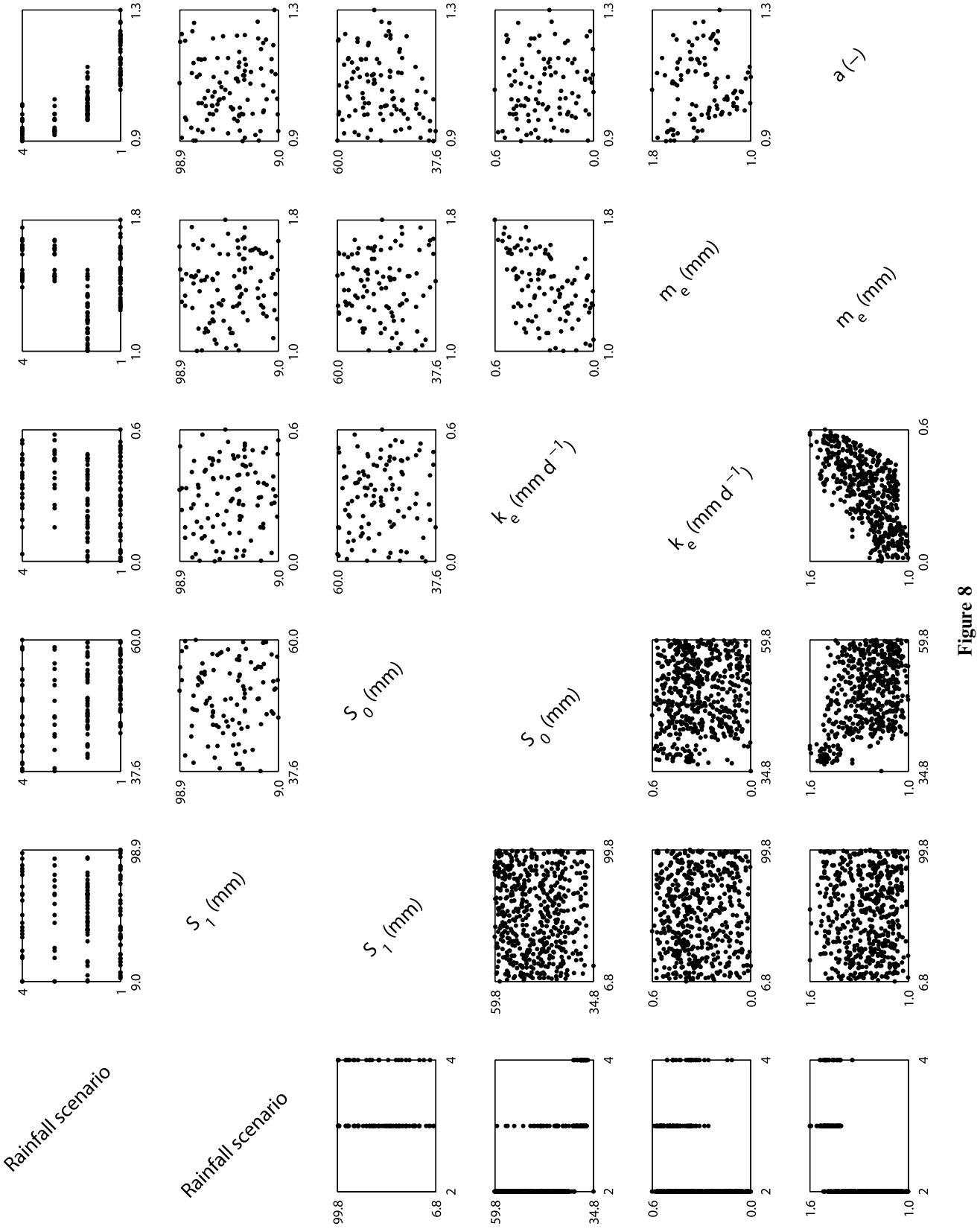

Figure 8

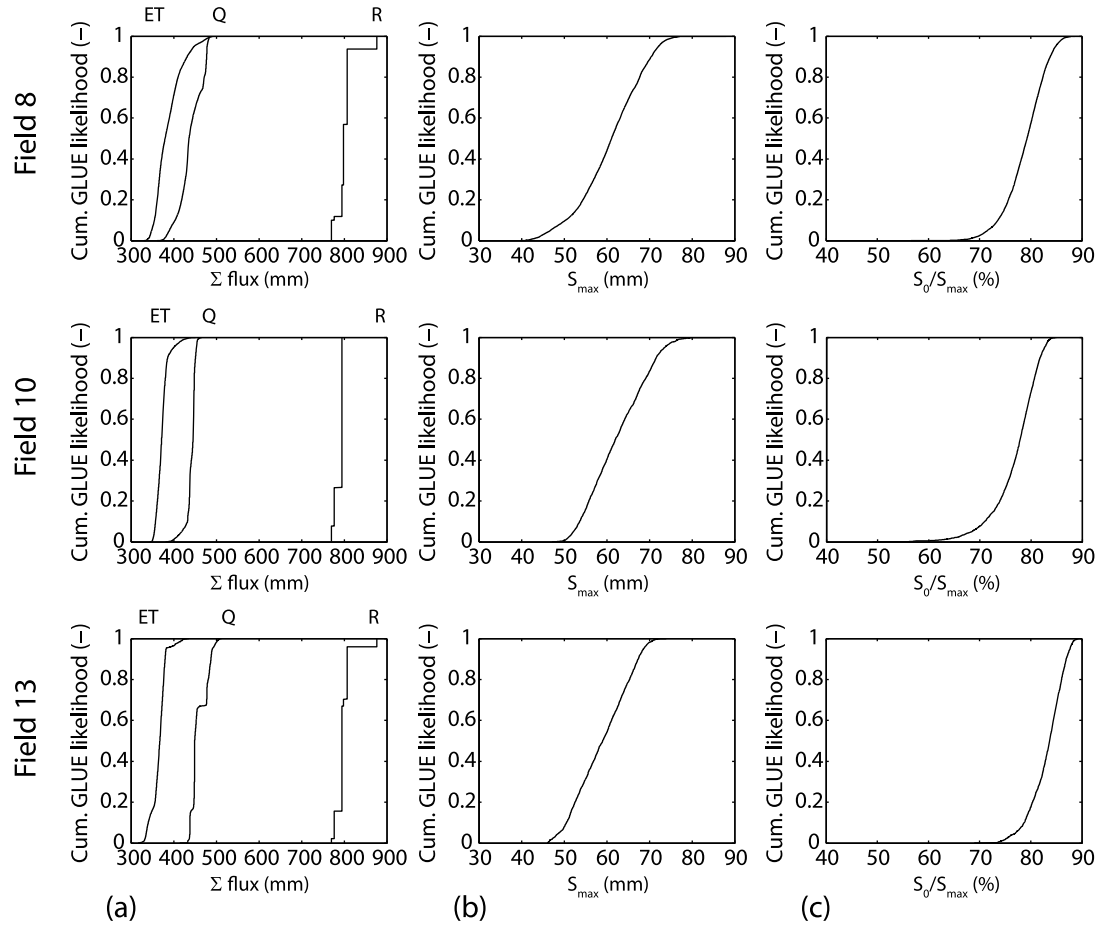

Figure 9. Cumulated GLUE likelihood distributions of simulation statistics for fields 8, 10 and 13: (a) simulated water balance expressed as total fluxes of rainfall (R), discharge (Q) and evapotranspiration (ET); (b) maximum simulated store (S_{max}); and (c) inactive store (S_0) expressed as percentage of S_{max}. All totals are summed over the 2006 Water Year (01/10/2005–30/09/2006).

tional purposes even if the data uncertainties are known well, means that some mismatch between simulations and observations has usually to be accepted on top of what is estimated as discharge measurement error. The limits of acceptability may not always be obvious and will depend on the intended use of the models, for diagnostics or different types of operational prediction, in which case the limits need to be defined post-hoc. This paper introduced a flexible methodology for doing so, based on time step–based performance measuring and performance aggregation over meaningful periods of the hydrograph. The limits of model acceptability were defined relative to the estimated discharge uncertainty intervals so that they served as indicators of model structural error (and model input error). More models should be evaluated in this way so that a series of benchmarks can build up which will help to appreciate the scale of model structural error for any given limits of acceptability that are expressed as multiples of measurement error.

[43] Rainfall input error was approached using rainfall scenarios in this study. The scenarios were found to be correlated with the resulting model parameter estimates, which indicates compensational effects between inputs and inferred model processes. This emphasizes the need for including input uncertainty in model evaluation to avoid rejecting behavioral models through biased inputs. The same can be implied for evapotranspiration uncertainty, which was not accounted for explicitly in this study. A quantification of evapotranspiration uncertainty would appear difficult to achieve beyond rough estimates in most cases due to the difficult task of measuring evapotranspiration in the first place. There is, consequently, a need for better scientific understanding of all observational uncertainties in hydrology through repeated experiments, novel measurement techniques and clustered instrumentation. Observational uncertainties should then be routinely incorporated into model diagnostic schemes to focus on the model structural error component and arrive, eventually, at a more realistic set of model parameters and structures as working hypotheses for the description of hydrological systems.

[44] It has to be recognized, however, that the study reported in this paper was computationally demanding and the data generated became increasingly awkward to handle. For those situations, existing model diagnostic tools were developed further in this paper to display and diagnose model results comprehensively. In order to decrease run time and storage space, it is suggested that the model rejection step be

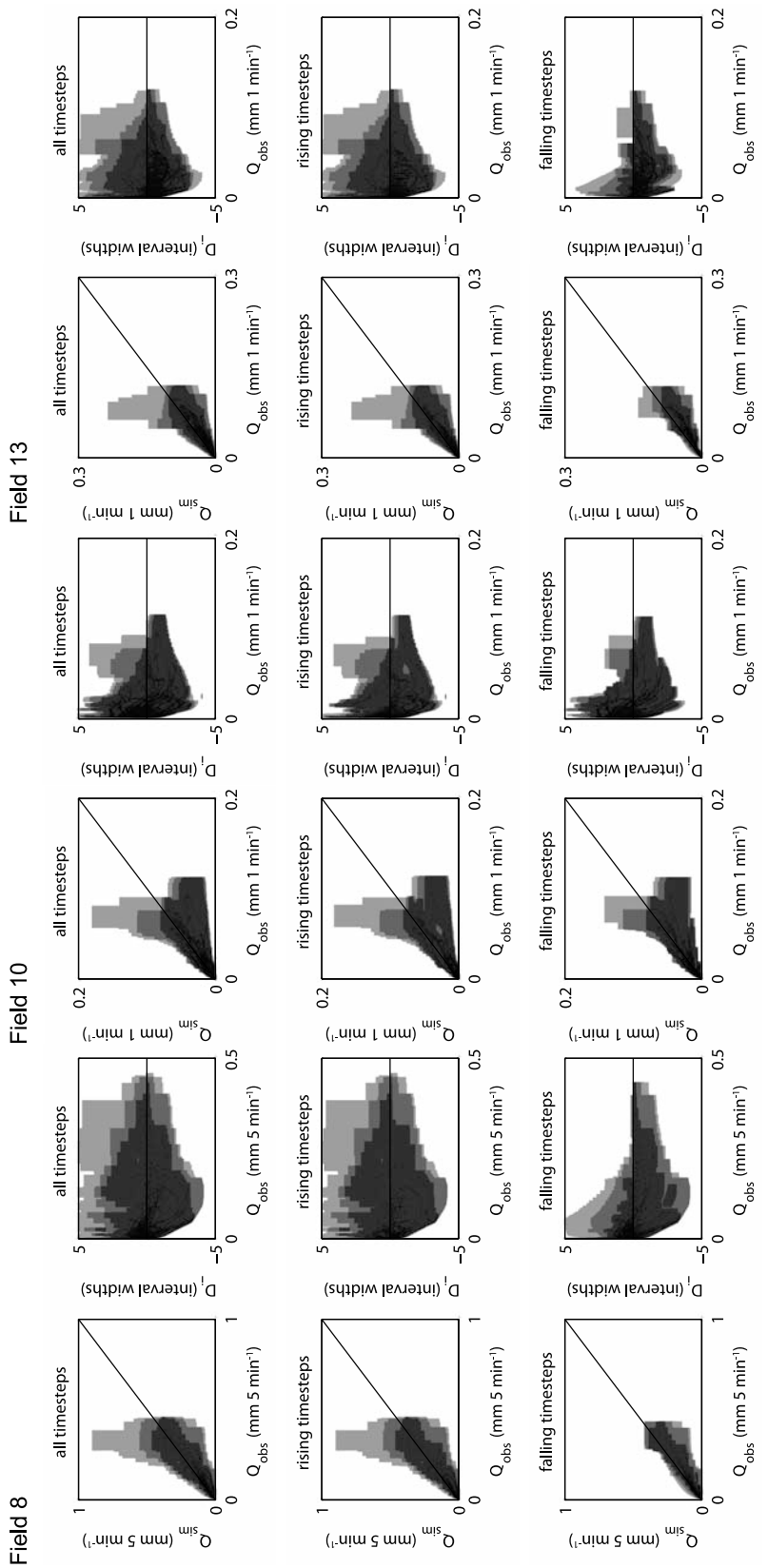

Figure 10. GLUE likelihood distributions of Q_{sim} and D_i, respectively, against Q_{obs} for fields 8, 10 and 13. The plots are repeated for the rising and recession (falling) time steps. The areas between the distribution percentiles min/max, 5th/95th and 25th/75th are shown in shades of grey of increasing intensity. The distribution medians are shown as points. A 1:1 line and a $D_i = 0$ line, respectively, is added for orientation.

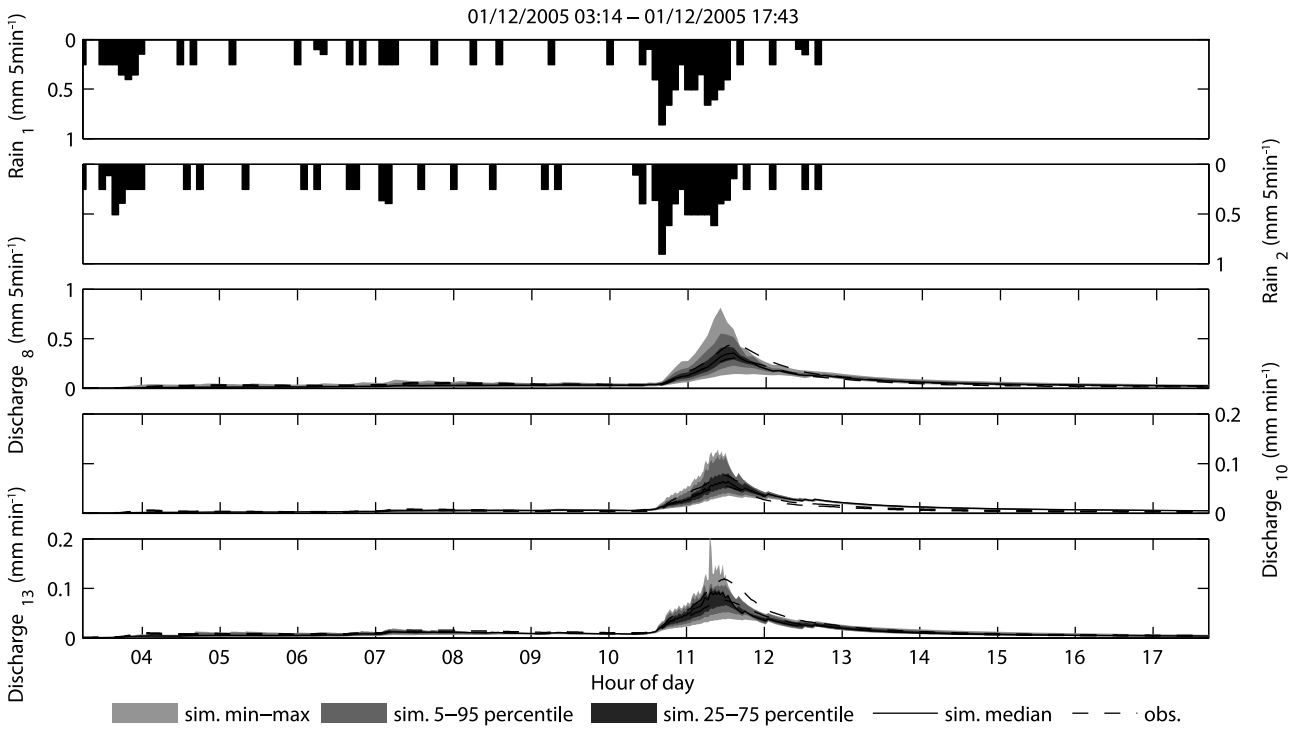

Figure 11. Observed rainfall event (gauges 1 and 2) and corresponding simulated (GLUE likelihood distribution) versus observed discharge for fields 8, 10 and 13.

simplified using pre-optimization [e.g., *Clark et al.*, 2008] in future applications, because only the optimum model performance is relevant for model rejection. This may also provide guidance on defining limits of model acceptability.

[45] As an example application, this paper marked the initial step in analyzing the hydrological behavior of a set of experimental field-scale lysimeters through model hypothesis testing. There were clear differences in model performance between fields which corresponded to real differences in hydrological behavior. For fields with events starting from low discharge levels, the single exponential or power law type of store reached its limit of applicability as an aggregated description of runoff generation at this small scale. The linear type of store and model structures without an inactive store were rejected. The bounded variants of stores caused unrealistic over-predictions through modeling overland flow as overspill routed to the field outlet in one 1–5 min time step. The alternative lumped simulation of overland flow and interflow seemed more realistic given that surface runoff may occur locally and may re-infiltrate before reaching the field boundary.

[46] All accepted model realizations were geared toward dissipating a large fraction of rainfall input by other means than discharge, resulting in simulations of actual evapotranspiration and inactive storage that were unrealistically large compared to field data estimates. It is hypothesized that the models compensated for a "leaking" of the fields, either through deep seepage despite the clay aquiclude, e.g., via macropores, or through the sides of the fields along the deep interceptor drains. In the spirit of model learning, additional field measurements should now test these hypotheses, while an improved model should include an additional loss term, e.g., a second outlet of the conceptual store. In addition, explicit flow routing formulations should be tested to address the identified timing issues.

Appendix A

[47] Stage-discharge uncertainty was estimated using the following algorithm, adapted from the idea of a fuzzy rating curve [*Pappenberger et al.*, 2006]:

[48] 1. The experiments carried out at the two weirs, of the same design, were evaluated separately to allow for differences in the ratings of the structures that may exist.

[49] 2. The estimated error intervals of each measurement were visualized as data boxes in the stage-discharge space (Figure 2). The boxes of repeated measurements were joined resulting in one data box per stage increment. This allowed for the possibility of measurement errors being estimated too small, in which case they were adjusted based on the variability of repeats.

[50] 3. The flexible and widely used power law $Q = a\,(h + b)^c$ with discharge Q, stage h and parameters a, b and c was chosen as the rating equation. This choice reflects the defined nature of the weirs where this equation has some physical justification [*Chow*, 1959], yet no prior assumptions about the parameters were made. The parameter b accounts for errors in stage at zero discharge (accuracy of stage measurement).

[51] 4. The uncertainty envelope for the stage-discharge relationship based on the chosen rating equation was calculated semi-analytically as follows:

[52] (i) Iterate through all possible combinations of two data boxes. For each combination, iterate through two nested loops of the four corners of each of the two boxes. Iterate through a final nested loop of the two limits of the stage

interval at zero discharge ([−2; 2]; Table 2) and take these in turn as parameter b.

[53] (ii) With b defined, each iteration yields two values of Q and h and thus a system of two rating equations with two unknowns a and c. Calculate those analytically. Reject complex solutions for small h.

[54] (iii) Keep this realization of parameters if the resulting rating curve intersects all remaining data boxes.

[55] (iv) The minima and maxima of these rating curve realizations are an accurate representation of the envelope, i.e., the intervals of model parameters and the intervals of Q for given h. Despite a theoretical derivation, the accuracy of the algorithm was confirmed through random Monte Carlo sampling of the rating curve parameters.

[56] 5. For the two weirs which experiments were conducted for, the corresponding uncertainty envelopes were used. For all other weirs, both envelopes were combined into one to reflect larger uncertainties when no experiment was conducted yet acknowledging the expected similar behavior of similar structures.

[57] **Acknowledgments.** The research reported in this paper was undertaken under project PE0120 funded by UK Defra. North Wyke Research is a UK BBSRC funded research institute. Additional funding came from the UK NERC Flood Risk from Extreme Events (FREE) programme (grant NE/E002242/1) and the UK Research Councils Rural Economy and Land Use (RELU) programme (grant RES-229-25-0009-A). We thank Neil McIntyre (Imperial College London, UK), Giuliano Di Baldassarre (University of Bristol, UK), Martyn Clark (NIWA, New Zealand) and Alberto Montanari (University of Bologna, Italy) for their constructive comments on the manuscript, and Keith Beven (Lancaster University, UK) for his comments on an earlier version of this paper.

References

Abrahart, R. J., and L. See (2002), Multi-model data fusion for river flow forecasting: An evaluation of six alternative methods based on two contrasting catchments, *Hydrol. Earth Syst. Sci.*, *6*(4), 655–670.

Ajami, N. K., Q. Y. Duan, X. G. Gao, and S. Sorooshian (2006), Multimodel combination techniques for analysis of hydrological simulations: Application to distributed model intercomparison project results, *J. Hydrometeorol.*, *7*(4), 755–768.

Ajami, N. K., Q. Y. Duan, and S. Sorooshian (2007), An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction, *Water Resour. Res.*, *43*, W01403, doi:10.1029/2005WR004745.

Armstrong, A. C., and E. A. Garwood (1991), Hydrological consequences of artificial drainage of grassland, *Hydrol. Processes*, *5*(2), 157–174.

Avery, B. W. (1980), *Soil Classification for England and Wales*, *Soil Surv. Tech. Monogr.*, vol. 14, Rothamsted Exp. Stn., Harpenden, U. K.

Beck, M. B. (1987), Water quality modeling: A review of the analysis of uncertainty, *Water Resour. Res.*, *23*(8), 1393–1442.

Beven, K. (2005), On the concept of model structural error, *Water Sci. Technol.*, *52*(6), 167–175.

Beven, K. (2006), A manifesto for the equifinality thesis, *J. Hydrol.*, *320*(1–2), 18–36.

Beven, K. J., and A. Binley (1992), The future of distributed models—Model calibration and uncertainty prediction, *Hydrol. Processes*, *6*(3), 279–298.

Boyle, D. P., H. V. Gupta, and S. Sorooshian (2000), Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, *Water Resour. Res.*, *36*(12), 3663–3674.

Butts, M. B., J. T. Payne, M. Kristensen, and H. Madsen (2004), An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation, *J. Hydrol.*, *298*(1–4), 242–266.

Chow, V. T. (1959), *Open-Channel hydraulics*, McGraw-Hill, New York.

Clark, M. P., A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener, and L. E. Hay (2008), Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, *Water Resour. Res.*, *44*, W00B02, doi:10.1029/2007WR006735. [Printed *45*(12), 2009]

Duan, Q. Y., N. K. Ajami, X. G. Gao, and S. Sorooshian (2007), Multimodel ensemble hydrologic prediction using Bayesian model averaging, *Adv. Water Resour.*, *30*(5), 1371–1386.

Freer, J. E., K. J. Beven, and B. Ambroise (1996), Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the glue approach, *Water Resour. Res.*, *32*(7), 2161–2173.

Freer, J. E., K. J. Beven, and N. E. Peters (2003), Multivariate seasonal period model rejection within the Generalized Likelihood Uncertainty Estimation procedure, in *Calibration of Watershed Models*, *Water Sci. Appl.*, vol. 6, edited by Q. Duan et al., pp. 69–88, AGU, Washington, D. C.

Freer, J. E., H. McMillan, J. J. McDonnell, and K. J. Beven (2004), Constraining dynamic topmodel responses for imprecise water table information using fuzzy rule based performance measures, *J. Hydrol.*, *291*(3–4), 254–277.

Georgakakos, K. P., D. J. Seo, H. Gupta, J. Schaake, and M. B. Butts (2004), Towards the characterization of streamflow simulation uncertainty through multimodel ensembles, *J. Hydrol.*, *298*(1–4), 222–241.

Harmel, R. D., and P. K. Smith (2007), Consideration of measurement uncertainty in the evaluation of goodness-of-fit in hydrologic and water quality modeling, *J. Hydrol.*, *337*(3–4), 326–336.

Hsu, K.-l., H. Moradkhani, and S. Sorooshian (2009), A sequential Bayesian approach for hydrologic model selection and prediction, *Water Resour. Res.*, *45*, W00B12, doi:10.1029/2008WR006824.

Huard, D., and A. Mailhot (2006), A Bayesian perspective on input uncertainty in model calibration: Application to hydrological model "abc," *Water Resour. Res.*, *42*, W07416, doi:10.1029/2005WR004661.

Huard, D., and A. Mailhot (2008), Calibration of hydrological model GR2M using Bayesian uncertainty analysis, *Water Resour. Res.*, *44*, W02424, doi:10.1029/2007WR005949.

Intermap Technologies (2007), NEXTMap Britain: Digital terrain mapping of the UK, NERC Earth Observ. Data Cent., Didcot, U. K. (Available at http://badc.nerc.ac.uk/view/neodc.nerc.ac.uk_ATOM_dataen_11658383444211836)

Jothityangkoon, C., M. Sivapalan, and D. L. Farmer (2001), Process controls of water balance variability in a large semi-arid catchment: Downward approach to hydrological model development, *J. Hydrol.*, *254*(1–4), 174–198.

Kavetski, D. N., S. W. Franks, and G. Kuczera (2003), Confronting input uncertainty in environmental modelling, in *Calibration of Watershed Models*, *Water Sci. Appl.*, vol. 6, edited by Q. Duan et al., pp. 49–68, AGU, Washington, D. C.

Kavetski, D., G. Kuczera, and S. W. Franks (2006a), Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, *Water Resour. Res.*, *42*, W03407, doi:10.1029/2005WR004368.

Kavetski, D., G. Kuczera, and S. W. Franks (2006b), Bayesian analysis of input uncertainty in hydrological modeling: 2. Application, *Water Resour. Res.*, *42*(3), W03408, doi:10.1029/2005WR004376.

Kavetski, D., G. Kuczera, and S. W. Franks (2006c), Calibration of conceptual hydrological models revisited: 1. Overcoming numerical artefacts, *J. Hydrol.*, *320*(1–2), 173–186.

Kennedy, M. C., and A. O'Hagan (2001), Bayesian calibration of computer models, *J. R. Stat. Soc., Ser. B*, *63*, 425–450.

Kirkby, M. (1975), Hydrograph modelling strategies, in *Processes in Physical and Human Geography*, edited by R. Peel, M. Chisholm, and P. Haggett, pp. 69–90, Heinemann, London.

Klemeš, V. (1983), Conceptualization and scale in hydrology, *J. Hydrol.*, *65*(1–3), 1–23.

Kuczera, G., D. Kavetski, S. Franks, and M. Thyer (2006), Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters, *J. Hydrol.*, *331*(1–2), 161–177.

Liu, Y., J. Freer, K. Beven, and P. Matgen (2009), Towards a limits of acceptability approach to the calibration of hydrological models: Extending observation error, *J. Hydrol.*, *367*(1–2), 93–103.

Marshall, L., A. Sharma, and D. Nott (2006), Modeling the catchment via mixtures: Issues of model specification and validation, *Water Resour. Res.*, *42*, W11409, doi:10.1029/2005WR004613.

Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models, part I—A discussion of principles, *J. Hydrol.*, *10*(3), 282–290.

Neuman, S. P. (2003), Maximum likelihood Bayesian averaging of uncertain model predictions, *Stochastic Environ. Res. Risk Assess.*, *17*(5), 291–305.

Pappenberger, F., K. J. Beven, N. M. Hunter, P. D. Bates, B. T. Gouweleeuw, J. Thielen, and A. P. J. de Roo (2005), Cascading model uncertainty from

medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European flood forecasting system (EFFS), *Hydrol. Earth Syst. Sci.*, *9*(4), 381–393.

Pappenberger, F., P. Matgen, K. J. Beven, J. B. Henry, L. Pfister, and P. Fraipont de (2006), Influence of uncertain boundary conditions and model structure on flood inundation predictions, *Adv. Water Resour.*, *29*(10), 1430–1449.

Poeter, E., and D. Anderson (2005), Multimodel ranking and inference in ground water modeling, *Ground Water*, *43*(4), 597–605.

Priestley, C. H. B., and R. J. Taylor (1972), On the assessment of surface heat flux and evaporation using large-scale parameters, *Mon. Weather Rev.*, *100*(2), 81–92.

Richards, L. A. (1931), Capillary conduction of liquids in porous mediums, *Physics*, *1*(5), 318–333.

See, L., and R. J. Abrahart (2001), Multi-model data fusion for hydrological forecasting, *Comput. Geosci.*, *27*(8), 987–994.

See, L., and S. Openshaw (2000), A hybrid multi-model approach to river level forecasting, *Hydrol. Sci. J.*, *45*(4), 523–536.

Shamseldin, A. Y., and K. M. O'Connor (1999), A real-time combination method for the outputs of different rainfall-runoff models, *Hydrol. Sci. J.*, *44*(6), 895–912.

Shamseldin, A. Y., K. M. Oconnor, and G. C. Liang (1997), Methods for combining the outputs of different rainfall-runoff models, *J. Hydrol.*, *197*(1–4), 203–229.

Sivapalan, M., and P. C. Young (2005), Downward approach to hydrological model development, in *Encyclopedia of Hydrological Sciences*, vol. 3, edited by M. G. Anderson and J. J. McDonnell, pp. 2081–2098, John Wiley, Chichester, U. K.

Smith, P. J., K. J. Beven, and J. A. Tawn (2008), Detection of structural inadequacy in process-based hydrological models: A particle-filtering approach, *Water Resour. Res.*, *44*, W01410, doi:10.1029/2006WR005205.

Thornley, J. H. M., and I. R. Johnson (1990), *Plant and Crop Modelling: A Mathematical Approach to Plant and Crop Physiology*, Oxford Univ. Press, New York.

Vache, K. B., and J. J. McDonnell (2006), A process-based rejectionist framework for evaluating catchment runoff model structure, *Water Resour. Res.*, *42*, W02409, doi:10.1029/2005WR004247.

Vereecken, H., J. A. Huisman, H. Bogena, J. Vanderborght, J. A. Vrugt, and J. W. Hopmans (2008), On the value of soil moisture measurements in vadose zone hydrology: A review, *Water Resour. Res.*, *44*, W00D06, doi:10.1029/2008WR006829.

Vrugt, J. A., and B. A. Robinson (2007), Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging, *Water Resour. Res.*, *43*, W01411, doi:10.1029/2005WR004838.

Vrugt, J. A., H. V. Gupta, W. Bouten, and S. Sorooshian (2003), A shuffled complex evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters, *Water Resour. Res.*, *39*(8), 1201, doi:10.1029/2002WR001642.

Vrugt, J. A., C. G. H. Diks, H. V. Gupta, W. Bouten, and J. M. Verstraten (2005), Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation, *Water Resour. Res.*, *41*, W01017, doi:10.1029/2004WR003059.

Vrugt, J. A., M. P. Clark, C. G. H. Diks, Q. Duan, and B. A. Robinson (2006), Multi-objective calibration of forecast ensembles using Bayesian model averaging, *Geophys. Res. Lett.*, *33*, L19817, doi:10.1029/2006GL027126.

Vrugt, J. A., C. J. F. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson (2008), Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, *Water Resour. Res.*, *44*, W00B09, doi:10.1029/2007WR006720.

Wagener, T., and J. Kollat (2007), Numerical and visual evaluation of hydrological and environmental models using the Monte Carlo analysis toolbox, *Environ. Modell. Software*, *22*(7), 1021–1033.

Wagener, T., D. P. Boyle, M. J. Lees, H. S. Wheater, H. V. Gupta, and S. Sorooshian (2001), A framework for development and application of hydrological models, *Hydrol. Earth Syst. Sci.*, *5*(1), 13–26.

Wagener, T., N. McIntyre, M. J. Lees, H. S. Wheater, and H. V. Gupta (2003), Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis, *Hydrol. Processes*, *17*(2), 455–476.

Xiong, L. H., A. Y. Shamseldin, and K. M. O'Connor (2001), A non-linear combination of the forecasts of rainfall-runoff models by the first-order Takagi-Sugeno fuzzy system, *J. Hydrol.*, *245*(1–4), 196–217.

Ye, M., S. P. Neuman, and P. D. Meyer (2004), Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff, *Water Resour. Res.*, *40*, W05113, doi:10.1029/2003WR002557.

Young, P. C., and K. J. Beven (1994), Data-based mechanistic modeling and the rainfall-flow nonlinearity, *Environmetrics*, *5*(3), 335–363.

Younger, P. M., J. E. Freer, and K. J. Beven (2009), Detecting the effects of spatial variability of rainfall on hydrological modelling within an uncertainty analysis framework, *Hydrol. Processes*, *23*(14), 1988–2003.

G. S. Bilotta, School of Environment and Technology, University of Brighton, Lewes Road, Brighton BN2 4GJ, UK.

R. E. Brazier, Department of Geography, University of Exeter, Exeter EX4 4RJ, UK.

P. Butler and C. J. A. Macleod, Cross Institute Programme for Sustainable Soil Function, North Wyke Research, Okehampton EX20 2SB, UK.

J. Freer, School of Geographical Sciences, University of Bristol, Bristol BS8 1SS, UK.

P. M. Haygarth and J. N. Quinton, Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK.

T. Krueger, School of Environmental Sciences, University of East Anglia, Norwich NR4 7TJ, UK. (t.krueger@uea.ac.uk)